Utilización de las máquinas con vectores de soporte para regresión: m^2 de construcción en Bogotá

Using support vector machines for regression: m^2 building in Bogotá

Abraham Gómez Morales[1], M.Sc. & Germán Hernández[2], PhD.
1. Carrera de Ingeniería de Sistemas, Universidad Central de Colombia
2. Departamento de Ingeniería de Sistemas e Industrial, Universidad Nacional de Colombia, Sede Bogotá
agomezm@ucentral.edu.co, gjhernandezp@unal.edu.co

Recibido para revisión 17 de Septiembre de 2008, aceptado 25 de Agosto de 2009, versión final 14 de Septiembre de 2009

Resumen— **En este artículo se ofrece una idea global acerca del tema de máquinas con vectores de soporte y sus aplicaciones tanto para tareas de clasificación como para tareas de regresión, enfocándose en estas últimas con el objetivo de determinar cómo utilizarlas para el cálculo del metro cuadrado de construcción en Bogotá D.C.**

Palabras Clave— **Aprendizaje de máquinas y aprendizaje computacional, Redes neuronales artificiales, Algoritmos de inteligencia artificial, Aplicaciones de inteligencia artificial, Nuevas estrategias de optimización.**

Abstract— **This article provides an overview on support vector machines and their applications for classification and regression tasks, focusing on how to use them to forecast the value per square meter building in Bogotá D.C.**

Keywords— **Learning systems, ART neural networks, Algorithms, Artificial intelligence, Optimization methods.**

I. INTRODUCCIÓN

El presente artículo busca ofrecer una solución al problema de calcular de manera automática y objetiva, el valor del metro cuadrado de construcción de la ciudad de Bogotá a partir de la información contenida en los indicadores prediales de los predios ubicados en el perímetro urbano. Es preciso señalar que el cálculo del valor del metro cuadrado de construcción es un proceso que actualmente lo realizan avaluadores humanos entrenados e involucra un gran número de variables y otros factores subjetivos como la experiencia. El valor del metro cuadrado de construcción es uno de los dos componentes que determinan el avalúo catastral para los predios de la capital, el otro componente es el valor de metro cuadrado del terreno. Debido al crecimiento urbano de Bogotá y a que parte importante de los ingresos de la ciudad provienen del impuesto predial se han explorado diferentes alternativas para encontrar mecanismos automáticos que produzcan los avalúos catastrales de la manera más precisa posible.

Esta investigación es la continuación de dos trabajos previos en los que se desarrollaron modelos para estimar el valor del metro cuadrado de construcción: uno en el cuál se desarrolló un modelo econométrico y otro en el que se utilizaron redes neuronales artificiales. El primero de estos modelos era básicamente un modelo lineal restringido a los indicadores prediales más influyentes, este modelo aunque tenía una alta capacidad explicativa no ofrecía la precisión de estimación deseada. Por otro lado en el segundo modelo, basado en redes neuronales artificiales se obtuvo una precisión adecuada en el valor del metro cuadrado de construcción. En el presente trabajo se busca diseñar otro modelo de solución, esta vez basado en métodos de kernels y que tenga una precisión de estimación comparable a la del modelo basado en redes neuronales. Se exploraron varios modelos basados en regresión "ridge" con kernels polinomiales y regresión con vectores de soporte.

El mejor modelo obtenido, entre los explorados, fue un modelo basado en regresión ridge con un kernel polinomial de grado 5, el cual se ubica en un punto medio comparado con las soluciones anteriores.

En el artículo se encontrará un resumen acerca del problema, la información utilizada, sus antecedentes y una descripción de los trabajos previos, posteriormente se describirán los aspectos relevantes de los métodos de Kernels, las máquinas

con vector de soporte y su utilización en tareas de regresión. Después se presenta la organización de los experimentos realizados, los modelos encontrados y las comparaciones de manera gráfica y estadística de estos modelos. En la última parte se encuentran las conclusiones y el posible trabajo futuro que se deriva de esta investigación.

II. EXPLICACIÓN DEL PROBLEMA Y ANTECEDENTES DE SOLUCIÓN

A. El metro cuadrado de construcción y sus componentes

De acuerdo con [6] los ingresos del distrito capital tienen cuatro fuentes fundamentales: las transferencias nacionales, el impuesto de industria y comercio (ICA), el impuesto de automotores y el impuesto predial unificado, este último se calcula con base en el avaluó catastral el cual a su vez depende, para un predio, del valor del metro cuadrado de terreno y valor del metro cuadrado de construcción. El cálculo de los avalúos catastrales se ha convertido en un problema de grandes proporciones para el distrito, para darse una idea del tamaño del problema basta recordar que en el año 2007 fue necesario aplazar el cobro del impuesto predial basado en el avalúo catastral actualizado debido a que el distrito fue incapaz de explicar y justificar este avalúo para un gran volumen de predios que presentaron reclamaciones.

La entidad dentro del distrito que tiene la responsabilidad de calcular los avalúos catastrales de los inmuebles ubicados en el perímetro urbano de la ciudad de Bogotá, es el Departamento Administrativo de Catastro Distrital (DACD). En el año 2002 el DACD contrató a la Universidad Nacional para diseñar un modelo econométrico que encontrara relaciones entre los indicadores prediales y el valor del metro cuadrado de construcción. Para esto se utilizó como información 2627 registros con información de predios residenciales de Bogotá a 2002, cada uno de estos registros tiene 36 indicadores prediales que inciden en el valor del metro cuadrado de construcción. En 2004 buscando mejorar las capacidades de predicción, utilizando la misma información, se exploraron algunos modelos basados en redes neuronales, los cuales presentaron mejores resultados.

B. Indicadores prediales

Los indicadores prediales son los atributos de un predio o inmueble. En la información suministrada por el DACD los indicadores prediales utilizados son: Localidad, Estrato Socioeconómico, Tipo de Predio, Área de Terreno, Área Total Construida, Área Construida por Uso, Código de Uso, Participación del Terreno, Valor del metro cuadrado de Terreno, Valor Total Terreno, Puntaje, Valor Total Construcción por Uso, Valor Total Construcción del Predio, Avaluó Total, Integral Comercial, Año Construcción y Edad. Es importante aclarar el concepto de puntaje observado en la lista anterior, que corresponde a una calificación dada de manera subjetiva a la construcción según sus características físicas (Estructura, Acabados Principales, Baño y Cocina). Esta información está completa para los predios del distrito capital excepto para los predios nuevos que requerirían una visita para el cálculo del puntaje.

C. Avalúo predial (catastral)

El avalúo predial es el proceso de determinar el valor de los predios en función de sus características, esta correlacionado, pero puede ser diferente, al precio dado por el mercado inmobiliario que está determinado por las presiones de oferta y demanda. Depende de tres factores, a saber:

- Las características del predio (terreno y construcción), como lo son las áreas, los diferentes valores, la antigüedad, etc.
- La localización, por ejemplo, la zona de ubicación del inmueble, el estrato socioeconómico, etc.
- La situación macroeconómica, que involucra los tipos de interés y el crecimiento de la demanda entre otros.

Para el análisis y tabulación de la información de indicadores prediales se utiliza el modelo AVM (Automated Valuation Models), citado en [1] y que permite recoger los aspectos anteriores, por medio, de dos grandes modelos: hedónicos e indexación y además una técnica llamada de testigos.

Los modelos hedónicos estiman el valor de la vivienda a partir de sus atributos, que determinan la calidad del inmueble y que son menos sensibles a los cambios por demanda y oferta. En los valores que calculan estos modelos influyen de manera importante las características del entorno donde se ubican, además que para su correcto funcionamiento se necesita contar con los expertos necesarios. Entre las ventajas de esta opción se encuentra el hecho que consigue un alto grado de precisión en la estimación, debido a la discriminación que hace de los componentes del valor final, como desventaja se tiene que son limitados para explicar satisfactoriamente el mercado.

Los modelos de indexación son los más sencillos, ya que se construyen indicadores de evolución de mercados inmobiliarios locales y computan periódicamente la inflación inmobiliaria en zonas concretas. Su ventaja radica en la facilidad de explicación, y su desventaja es que no todas las clases de inmuebles, aún en la misma zona, experimentan el mismo crecimiento de precio.

En la técnica de testigos se seleccionan valoraciones recientes de inmuebles cercanos al inmueble que se pretende valorar (valores de referencia) y se extrae un indicador predial, por ejemplo, el valor del metro cuadrado de construcción en la zona y se pondera por los valores del mismo indicador predial en los otros inmuebles. Una de sus ventajas es la fácil comprensión de la técnica y que permite detectar cambios en mercados locales con bastante rapidez; su desventaja es que puede ser incapaz de valorar inmuebles en zonas poco densas o con pocas valoraciones.

D. Modelo econométrico (2003)

Este trabajo de investigación, referenciado en [10], lo hizo la oficina de extensión y asesoría y el área de consultoría del Departamento de Estadística de la Universidad Nacional de Colombia, sede Bogotá, y consistió en desarrollar un modelo econométrico para estimar el valor del metro cuadrado de vivienda en la zona urbana de Bogotá. Con miras a lograr esta estimación se utilizó un modelo hedónico que expresa la relación matemática entre el valor del metro cuadrado de construcción y una serie de indicadores prediales en cantidades que determinan dicho valor; entonces, a partir del modelo hedónico se hicieron dos análisis:

Un análisis estadístico descriptivo para predecir el comportamiento de las variables tanto individualmente como asociativamente, con el fin de verificar su influencia y poder de explicación, en términos de la variable "valor del metro cuadrado de construcción".

Se efectuó un análisis de componentes principales y correspondencias simples y múltiples, con la finalidad de encontrar posibles asociaciones entre las variables.

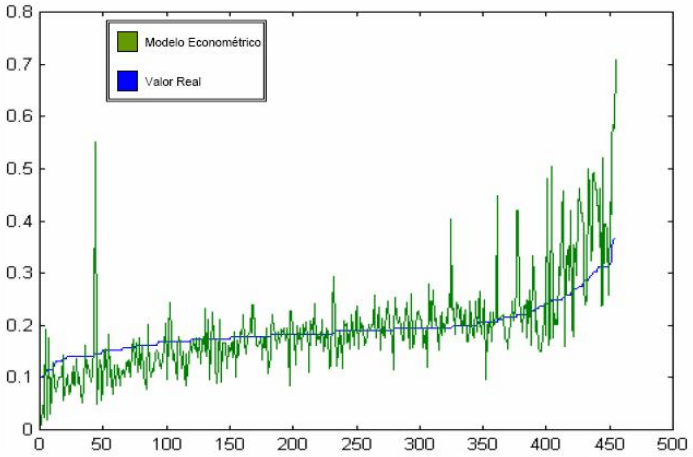
Después del anterior análisis se llegó a la ecuación número (1).

Resultados: Efectuada las mediciones por medio del modelo, se logran unos resultados, los cuales fueron ordenados por registros, de acuerdo al valor del metro cuadrado de construcción de manera ascendente, por medio de una fórmula de normalización que se indica en la ecuación (2).

Se evidencia, que la regresión debe tener una alta precisión, debido a que por la normalización cada centésima (0.01) de error representa $18.000 pesos (precios 2002) aproximadamente (Véase Gráfica 1).

$$y = \beta_0 + \beta_1 Puntaje + \beta_2 D6 + \beta_3 U38ED + \beta_4 U38TO + \beta_5 ValM^2 Ter + e \quad (1)$$

$$Valor\ m^2\ construcción(x) \in [3000, 2640000] \quad (2)$$

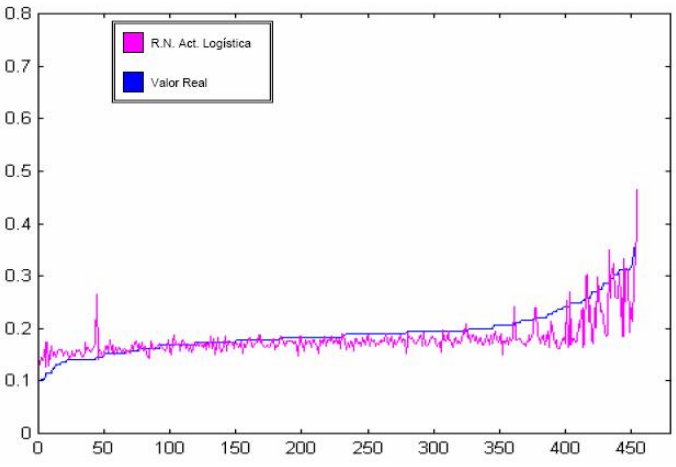


Gráfica 1. Valor real – modelo econométrico
Fuente: [2]

El error cuadrático de este modelo es de 1.9701 y el valor de ese error en Pesos Colombianos a precios de 2002 es de $3'546.180.oo. A raíz de los fundamentos teóricos con los que se desarrolló este modelo, su mayor deficiencia consiste en su incapacidad de adaptarse dinámicamente a los cambios de los valores de las variables, al tratarse de un sistema probabilístico de naturaleza estática.

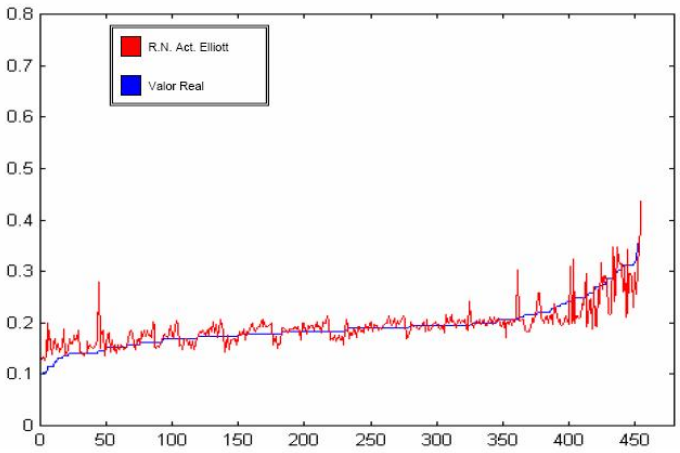
E. Modelo de redes neuronales (2004)

De acuerdo a los resultados obtenidos con el anterior modelo, fundamentándose en [9] y debido a la poca precisión ofrecida, se desarrolló por parte de los expertos de la Universidad Nacional una herramienta de software utilizando técnicas de Redes Neuronales Artificiales (RNA's) como herramientas de regresión y así poder calcular el valor del metro cuadrado de construcción en la ciudad de Bogotá, ya que utilizando este tipo de redes se puede generalizar el modelo econométrico, proporcionando así un avance significativo para la determinación del monto de los impuestos de renta en la ciudad y poder determinar el valor del metro cuadrado de construcción, todo lo anterior se reseña profundamente en [2]. Se vio entonces la necesidad de realizar otros experimentos con este nuevo método, con miras a lograr la exactitud deseada para el valor del metro cuadrado de construcción; analizando los resultados anteriores se pudo deducir que existen relaciones no lineales entre el metro cuadrado de construcción y los valores de los indicadores prediales y el objetivo era usar una RNA para modelar este tipo de relaciones.

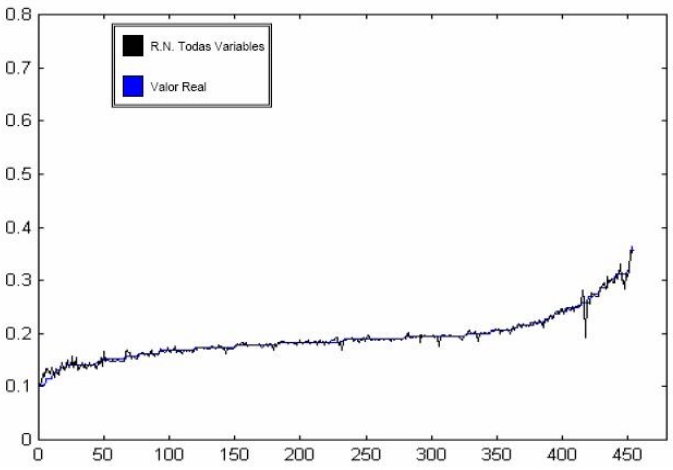


Gráfica 2. RNA Regresión No. 1 Función de Actividad Logística
Fuente: [2]

Para los experimentos se utilizaron los mismos 2627 registros que con el modelo econométrico, trabajándolos en proporción de 75% para el proceso de entrenamiento y 25% de los datos para pruebas. Se utilizó una RNA feedforward con el algoritmo de entrenamiento de retro propagación y se entrenó buscando que el error para todos los datos del conjunto de entrenamiento se minimizara en una milésima (0.001), lo que equivaldría a un error de $1.800. La herramienta de software con la que se diseñó la red neuronal fue JNNS y se hicieron dos regresiones:



Gráfica 3. RNA Regresión No. 2 Función de Actividad de Elliot
Fuente: [2]



Gráfica 4. RNA Regresión No. 2 Todos los Indicadores Prediales
Fuente: [2]

• La primera con los mismos indicadores prediales que el modelo econométrico, es decir, estrato socioeconómico, tipo de predio, valor del metro cuadrado de terreno, código de uso, puntaje total, valor del metro cuadrado de construcción y año de construcción y utilizando la función de activación logística y la función de activación de elliot.

• Utilizando el total de indicadores prediales, que son: puntaje, localidad, estrato socioeconómico, tipo de predio, área del terreno, valor del m2 del terreno, área total construida, código de uso, área construida por uso, participación del terreno en el uso, construcción por uso, construcción por predio, avaluó total, integral comercial y edad.

Tabla 1. Error Cuadrático 1° Regresión Vs. Econométrico

Comparativo de Errores	Econométrico	RNA función de activación de Elliott	RNA función de activación Logística
Error Cuadrático	1,9701	0,2633	0,4174
Error en Dinero	$3'546.180.oo	$ 473.940.oo	$ 751.320.oo

Fuente: autor

Tabla 2. Error Cuadrático 2° Regresión

Comparativo de Errores	Red Neuronal Usando Todos las Indicadores
Error Cuadrático	0.0160
Error en Dinero	$ 28.800.oo

Fuente: autor

Resultados: Efectuados los respectivos experimentos se encontraron unos resultados superiores a los resultados anteriores, los cuáles se ordenaron de igual forma que los del econométrico, es decir, de acuerdo al valor del metro cuadrado de construcción y de manera ascendente, utilizando la misma forma de normalización. Comparando los resultados de la primera regresión a nivel de gráficos (Véase Gráficas 2 y 3) se observan comportamientos parecidos del valor de metro cuadrado, sin embargo, a nivel de error cuadrático se observa (Véase Tabla 1) una marcada diferencia entre lo modelos obtenidos, favoreciéndose el modelo logrado con la función de activación de elliot, que permite un error de $473.940.

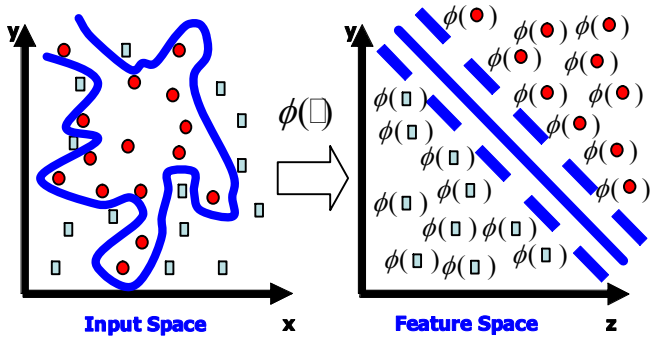
En la segunda regresión (Véase Gráfica 4) se observa un comportamiento totalmente superior a los anteriores ya que tenemos un error cuadrático medio muy pequeño (Véase Tabla 2) lo que favorece uno de los objetivos, en cuanto es una estimación bastante precisa que consigue un error de $28.800.

III. TEORÍA DE MÁQUINAS CON VECTORES DE SOPORTE

A. Conceptos

Las máquinas basadas en kernels, en particular las máquinas con vectores de soporte, según [4], [7] y [15], son técnicas de clasificación en las que los datos de entrada son transformados implícitamente a un espacio nuevo, generalmente de una dimensión superior, en donde es posible clasificarlos con hiperplanos de separación como se observa en la Gráfica 5. El espacio inicial es el indicado como espacio de entradas (input space) y el espacio que contiene los datos transformados es llamado espacio de características (feature space). A continuación, se presenta brevemente el aprendizaje basado en kernels y las máquinas con vectores de soporte relevantes al trabajo, para una visión más detallada se recomiendan los trabajos documentados en [3], [12] y [16].

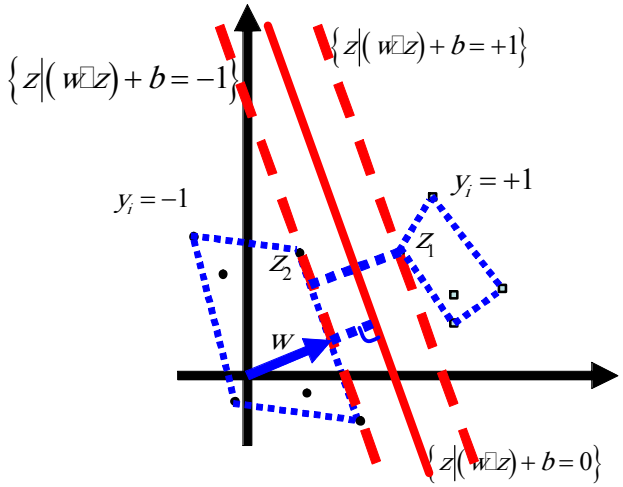
En estos métodos, la transformación no lineal de los datos de entrada no se calcula de manera explícita gracias al uso de una función kernel, que permite realizar todas las computaciones en el espacio de entrada, esto es llamado el kernel trick. En máquinas con vectores de soporte se busca el hiperplano óptimo de separación, el cual solo depende de unos pocos puntos llamados vectores de soporte. Ver [5], [11] y [17]. Gráfica 6.



Gráfica 5 Transformación de Espacios
Fuente: autor

B. Máquinas con vector de soporte para regresión

Para utilizar las máquinas con vector de soporte para tareas de regresión es necesario crear un concepto análogo al margen en el espacio de valores de salida $y \in \square$ mediante el uso de la función de costo (pérdida) $\varepsilon - insensitiva$:



Gráfica 6. Hiperplano óptimo
Fuente: [8]

$$|y - f(z, w)|_\varepsilon = \max\{0, |y - f(z, w)| - \varepsilon\} \quad (3)$$

Para valores de $\varepsilon \geq 0$ el problema de optimización del apartado de clasificación se transforma en:

$$\min(w, \varphi^{(*)}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\varphi_i + \varphi_i^*) \quad (4)$$

Sujeto a:

$$\begin{aligned} &(\langle w \cdot z_i \rangle + b) - y_i \leq \varepsilon + \varphi_i \\ &y_i - (\langle w \cdot z_i \rangle + b) \leq \varepsilon + \varphi_i^*, \varphi_i, \varphi_i^* \geq 0 \end{aligned} \quad (5)$$

con $i = 1, \ldots, \ell$

Al introducir los multiplicadores de Lagrange:

$$W(\alpha, \alpha^*) = \varepsilon\sum_{i=1}^{\ell}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)y_i + \frac{1}{2}\sum_{i,j=1}^{\ell}(\alpha_i^* - \alpha_i)\cdot k(z_i, z_j)\cdot(\alpha_j^* - \alpha_j) \quad (6)$$

Sujeto a:

$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, \ell \text{ y } \sum_{i=1}^{\ell}(\alpha_i - \alpha_i^*) = 0$$

Entonces la regresión toma la forma:

$$f(z) = \sum_{i=1}^{SV}(\alpha_i^* - \alpha_i)\cdot k(z_i, z) + b \quad (7)$$

y b se calcula nuevamente utilizando las condiciones de Karush-Kuhn-Tucker:

$$\alpha_i \cdot [y_i \cdot (\langle w, z \rangle + b) - 1] = 0 \quad (8)$$

con $i = 1, \ldots, \ell$

IV. EXPERIMENTACIÓN

En esta investigación se realizaron cuatro experimentos, procediendo de la siguiente manera: el conjunto de datos de entrada es de 2626 registros, de los cuales se tomaron 2100 de entrenamiento (80%) y 526 de prueba (20%). Se utilizó la función kernel polinomial, ya que la teoría apoya que en alguna dimensión superior, a lo más en la n+1, es posible pasar por todos los puntos, además se utilizó k-fold cross validation con valores k=5 (5 grupos de 420) y k=10 (10 grupos de 210). El modelo actual se encuentra entre dos tendencias:

- No es muy preciso pero generalizable.
- Es muy preciso pero no generalizable.

Se utilizaron los parámetros de regularización d, C, λ y ν y con valores así:

- El valor d se maneja a través de la función kernel, ya que es el grado del polinomio, con valores desde 2 hasta 8, por problemas de recurso de hardware.
- El valor C se encuentra entre 1 y 20, ya que a partir de este valor las pruebas de predicción de valores nuevos empezaban a tener errores.
- El parámetro λ entre 0.005 y 2 por problemas de computación (KRR).

Para los valores de se trabajó entre 0.5 y 0.8.

Los experimentos λ se hicieron con regresión ridge, SVR y -SVR. El método de regresión ridge es el más cercano a una regresión, cuando el valor es cercano a cero, por eso es el primer método que se utilizó.

A. Experimento Uno

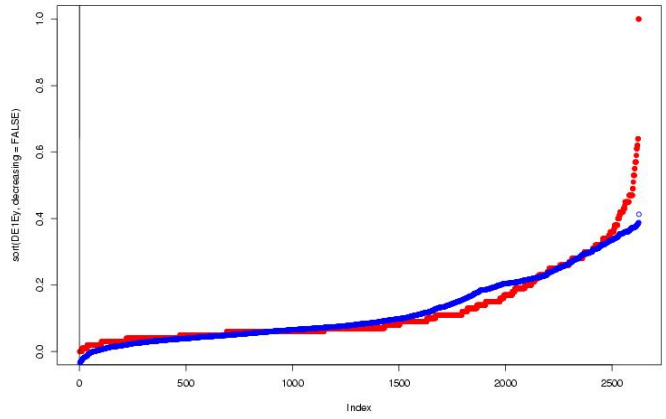
Que involucra los mismos indicadores prediales del modelo econométrico, los cuáles son: el estrato socioeconómico, el tipo de predio, el código de uso, el valor del metro cuadrado del terreno, el puntaje, y el año de construcción. El objetivo del experimento es determinar si el modelo con SVR presenta una mejor solución que el modelo econométrico, el de RNA con función de activación logística ó el de RNA con función de activación de Elliot, a nivel de los errores.

Resultados: El mejor resultado es el mostrado en la Gráfica 7 (Valor Real Color Rojo) que utiliza un modelo de regresión ridge con función kernel de grado de polinomio 5 y valor λ = 0.005. Este modelo presenta un error de 0.226267 y se compara con los modelos anteriores en la Tabla 3.

Tabla 3. Experimento uno – Tablas de errores

Comparativo de Errores	Modelo Econométrico	R.N.A. con función de activación de Elliott	R.N.A. con función de activación de Logística	KRR
Error Cuadrático	1,9701	0,2633	0,4174	0.226267
Error en Dinero	$3'546.180.oo	$ 473.940.oo	$ 751.320.oo	$ 405.857.oo

Fuente: autor

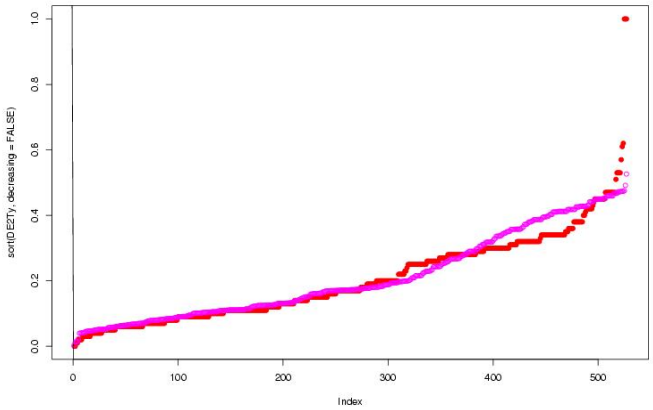


Gráfica 7. Experimento uno
Fuente: autor

B. Experimento Dos

Este es un experimento que utiliza un conjunto reducido de indicadores prediales, se basa en un modelo con RNA presentado por Quang y Grudnitski que reflejaba una relación entre el estrato socioeconómico, el valor del metro cuadrado del terreno, el puntaje y la edad, para la ciudad de San Diego (USA), para una mayor profundidad en este tema se recomienda leer [13] y [14]. El objetivo es comprobar si esa misma relación se mantiene para la ciudad de Bogotá.

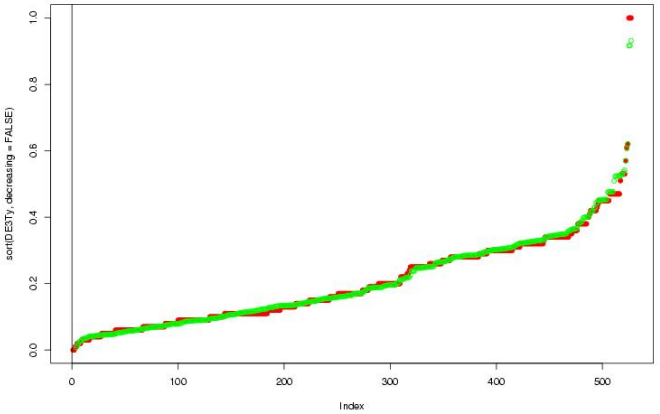
Resultados: El mejor resultado es el mostrado en la Gráfica 8 (Valor Real Color Rojo). Se utilizó un modelo de SVR con función kernel de grado de polinomio 4, parámetro C=2, parámetro ε = 0.001, y k-fold cross validation de valor 5. Los resultados de este modelo presentan un error de 0.3000834 pero no tienen contra que compararse.



Gráfica 8. Experimento dos
Fuente: autor

C. Experimento Tres

Este experimento utiliza el total de indicadores prediales: la localidad, el estrato socioeconómico, el tipo de predio, el área de terreno, el área total construida, el área de construcción por uso, el código de uso, la participación del terreno, el valor del metro cuadrado del terreno, el puntaje, el valor total de la construcción por uso, el valor total de construcción del predio, el avaluó total, el integral comercial y la edad. El Objetivo es determinar si el modelo con SVR presenta una mejor solución que el modelo de RNA, a nivel de los errores.



Gráfica 9. Experimento tres
Fuente: autor

Resultados: El mejor de los resultados es el mostrado en la Gráfica 9 (*Válor Real Color Rojo*) que utiliza un modelo SVR con función kernel de grado de polinomio 5, parámetro C=10, parámetro ε =0.01, y *k-fold cross validation* de valor 10. Se presentó un error de 0.1000834 que se compara con los modelos anteriores en la Tabla 4.

Tabla 4. Experimento tres – Tablas de errores

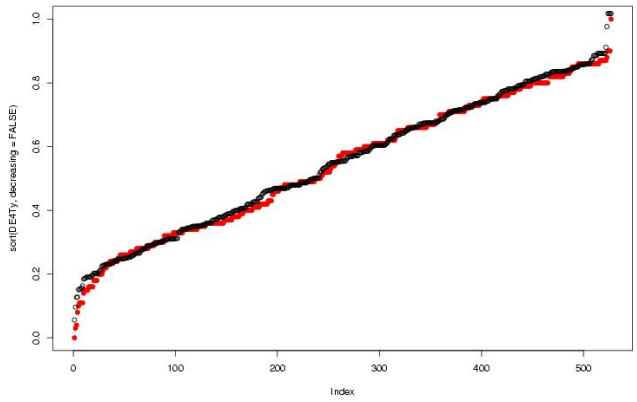
Comparativo de Errores	R.N.A. con todos los indicadores prediales	SVR
Error Cuadrático	0.0160	0.1000834
Error en Dinero	$ 28.700.oo	$ 179.521.oo

Fuente: autor

D. Experimento Cuatro

El objetivo de este experimento es determinar si indicador predial utilizado (puntaje) tiene un patrón con respecto a sus indicadores de origen. Los indicadores prediales involucrados son: la estructura de armazón, la estructura de muros, la estructura de la cubierta, la conservación de la estructura, los acabados de la fachada, los acabados de los cubrimientos, los acabados de los pisos, la conservación de los acabados, en cuanto a los baños, el tamaño, los enchapes, el mobiliario y la conservación de los mismos y en cuanto a la cocina, el tamaño, los enchapes, el mobiliario y la conservación de ésta.

Resultados: El mejor resultado lo aportó el modelo SVR con función kernel de grado de polinomio 4, parámetro C=10, parámetro ε =0.01, k-fold cross validation de valor 10 y un error medio cuadrático de 0.2002506. Véase la Gráfica 10 (Valor Real Color Rojo).



Gráfica 10. Experimento cuatro
Fuente: autor

V. CONCLUSIONES

Los métodos de regresión de las máquinas con vectores de soporte ofrecieron buenos resultados en comparación con los obtenidos con el modelo econométrico, la regresión lograda ganó en exactitud. De igual forma se comparó contra los resultados de la RNA que utiliza todos los indicadores prediales como entrada, aquí el resultado de la RNA es mejor que la del método con máquinas con vectores de soporte, sin embargo, se puede decir un poco más del proceso interiormente ya que tenemos nociones de qué se hace con el conjunto de datos de entrada, cosa que no es posible con las RNA's. En general, se puede afirmar que los modelos de regresión no lineales ofrecen mejores resultados que el modelo estadístico lineal (econométrico), esto verifica que el patrón encontrado es un patrón de relaciones no lineales entre los indicadores prediales y el valor del metro cuadrado de construcción y que no se trata de una función lineal, ya que esta hubiera sido fácilmente solucionada con el modelo econométrico. Existen algunos casos de puntos de entrenamiento anormales pero son inconsistencias en los registros del conjunto de datos de entrada, sin embargo, el error medio cuadrático está acorde con las perspectivas que se tenían. En general los errores no fueron muy grandes para las maquinas con vectores de soporte.

Los resultados apoyan la conclusión que es mejor entrenar el modelo con todos los indicadores prediales, lo más posible es que esto se deba a que al poseer más información de entrada se puede encontrar un mejor patrón de la fuente de datos. Es importante observar que con este modelo se puede obviar el análisis estadístico para seleccionar los indicadores prediales de entrada. Al ser la regresión ridge un método de máquinas con vectores de soporte para regresiones es más cercana al concepto de regresión, sin embargo, el cálculo de la matriz hace complejo intentar experimentos de este tipo con conjuntos grandes de datos, por esto, apoyándonos en la teoría y en los cuatro experimentos realizados, se concluye que las máquinas con vectores de soporte no se pueden usar con conjuntos de datos de entrada extensos, aunque no fue posible determinar qué significa el término extenso (numéricamente).

Las máquinas con vectores de soporte aprendieron información que estaba oculta en el conjunto de datos de entrada, concretamente los casos de los experimentos dos y cuatro. En cuanto al experimento dos (conjunto reducido de variables), se logró establecer una relación directa entre el estrato socioeconómico, el puntaje y la edad del inmueble con respecto a la estimación del valor del metro cuadrado de construcción. En el cuatro, anteriormente se asumía que el puntaje tenía una relación con sus componentes pero no había experimentos que apoyaran esta idea, ahora con los experimentos sabemos que es así, por lo que encontrar la relevancia entre estos datos constituye una garantía de credibilidad, pues en el actual modelo que utiliza el DACD esas relaciones se intuyen sin tener unos datos exactos que prueben esa relación.

El presente trabajo aplica la idea de hacer regresión con máquinas con vectores de soporte a un caso muy específico, no existen garantías que permitan saber si es posible aplicar esta idea sobre los datos de otra ciudad, es más, es posible

que existan otros indicadores prediales, que no fueron tenidos en cuenta en este trabajo. Por esto, y pese a la existencia de trabajos previos, los resultados son únicos y funcionales para cada caso en particular. En este sentido, una recomendación sería la realización de un texto que condense los diferentes resultados.

VI. TRABAJO FUTURO

Como trabajo futuro se propone mejorar el proceso de obtención de los datos in situ. De igual forma no se pudo observar como responderían las máquinas con vectores de soporte ante valores incoherentes presentes en los datos del conjunto de entrada, debido a que se realizó un trabajo de normalización diferente al que se hizo con las RNA's. En Bogotá, se necesita una herramienta que sirva de garantía y se ciña a la realidad económica de sus contribuyentes, razón por la cual el gobierno debe invertir constantemente en asesorías de esta clase, con miras a optimizar su servicio.

VII. REFERENCIAS BIBLIOGRÁFICAS

[1] Adair, A., McGreal, W., Computer Assisted Valuation of Residential Property. Review: The Real Estate Appraisal and Analyst. Winter Edition 1988. pp 18-21. USA. 1988

[2] Ávila, L., Robayo, V., Red Neuronal para Determinar el Valor del Metro Cuadrado de Construcción. Tesis. Universidad Nacional de Colombia. 2003

[3] Burges, C. A., Tutorial on Support Vector Machines for Pattern Recognition. Bell Laboratories, Lucent Technologies 1998

[4] Chen, Y., Wang, J. Z., Kernel Machines and Additive Fuzzy Systems: Classification and Function Approximation. The Pennsylvania State University. 2001

[5] Drucker, H., Burges, C., Kaufman, L., Smola ,A., Vapnik, V., Support Vector Regression Machines. Advances in Neural Information Processing Systems, The MIT Press, Cambridge, Massachusetts, 1997, page 155. 1996

[6] Estatuto Tributario de Santa Fe de Bogotá, D.C., Secretaría de Hacienda, 2000

[7] Gärtner, T. A., Survey of Kernels for Structured Data. University of Bonn. 2003

[8] Hearst, M., How to Implement Support Vector Machines. IEEE Intelligent Systems Page 18-28. July/August-1998

[9] Karakozova, O. A., Comparison between neural network and multiple regression approaches: An Application to Residential Valuation in Findland. Swedish School of Economics and Business Administration. 2000

[10] Oficina de Extensión y Asesoría de la Universidad Nacional de Colombia. Elaboración de Modelos Econométricos División de Actualización DACD. Facultad de Ciencias. Departamento de Estadística, Bogotá. 2002

[11] Pontil, M., Verry, A., Properties of Support Vector Machines. MIT AI Laboratory and Center for Biological and Computational Learning. 1997

[12] Pontil, M., Rifkin, R., Evgeniou, T., From Regression to Classification in Support Vector Machines. MIT AI Laboratory and Center for Biological and Computational Learning. 1998

[13] Quang, A., Grudnitski, G., A Neural Network Approach to Residential Property Appraisal. Review: The Real Estate Appraisal and Analyst. Volume 58. No. 3. pp 38-45. USA. Diciembre 1992

[14] Quang, A., Grudniski, G. A., Neural Network Analysis of the Effect of Age on Housing Values. Review: The Journal of Real Estate Research. Volume 8. No. 2. pp 253-264. USA. Primavera 1993

[15] Shawe-Taylor, J., Cristianini, N., Kernel Methods for Pattern Analysis. Cambridge Press, 2004.

[16] Smola, A., Schölkopf, B., Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning). MIT Press. 2001

[17] Vapnik, V., The Nature of Statistical Learning Theory. Springer. 2000